Journal of proteome research

LETTER

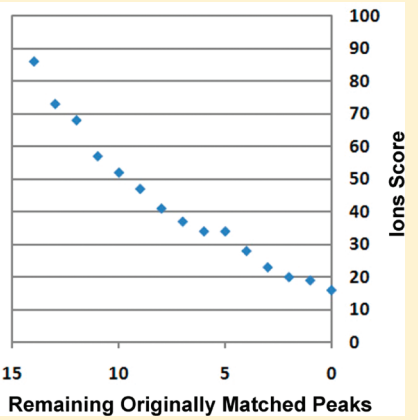
pubs.acs.org/jpr

The Problem with Peptide Presumption and Low Mascot Scoring

Bret Cooper*

Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, Maryland 20705, United States

ABSTRACT: Mascot, a database-search algorithm, is used to deduce an amino acid sequence from a peptide tandem mass spectrum. The magnitude of the Ions score associated with each peptide mostly reflects the extent of *b*−*y* ion matching in a collision-induced dissociation spectrum. Recently, several studies have reported peptides identified with abnormally low Ions scores. While a majority of the spectra in these studies may be correctly assigned, low-scoring spectra could lack discernible *b*−*y* ion fragments needed to clearly delineate a peptide sequence. It appears that low-scoring identification may be predicated primarily on judgmental parent ion mass accuracy and that justification to include such low-scoring peptides may be based on inaccurate false discovery rate modeling. It is likely that additional scientific experimentation is needed or appropriate methodologies adopted before substandard fragment ion matching can be considered proof of peptide identification.



KEYWORDS: Orbitrap, Mascot, mass accuracy, false discovery rate, reverse database

Over the last 2 years, numerous journals have published high-profile papers describing significant peptides identified with very low Mascot Ions scores. Soufi et al.[1] and Choudhary et al.[2] claimed peptides with scores as low as 10, Choudhary et al.[3] claimed peptides with scores as low as 5, Gnad et al.[4] accepted Ions scores as low as 1.9, and Olsen et al.[5] included peptides that scored 0.1. In all of these cases, 30−50% of the peptides were identified with scores less than 30. These authors possibly rationalize their inclusion of low-scoring peptides because they used (1) high-mass accuracy mass spectrometers (i.e., Orbitrap and/or FTICR) for parent ion mass (PIM) determination and (2) target-decoy database searching to control false discovery rates (FDRs). Here, I argue why these data sets are controversial.

First, let me define the Mascot Ions score. The Ions score reflects the number of observed MS/MS ions that match within a prescribed fragment ion mass tolerance any of the hypothetical MS/MS ions of a given peptide amino acid sequence. High scores are associated with a greater number of matches whereas a score of zero means no matches were made or that the matches that were made were not better than expected by chance.[6]

To illustrate the meaning of the Ions score with respect to peptide identification, a PIM for a molecule from *Salmonella enterica* was resolved in an Orbitrap and an MS/MS spectrum was generated from it in an LTQ linear ion trap with lower mass accuracy. As in the questioned studies, narrow ±10 ppm PIM tolerance (±0.01 Da for a 1000 Da molecule) but wider, ± 0.5 Da fragment ion mass tolerance parameters were used in Mascot searches to respectively reflect the mass accuracy of each analyzer. Mascot matched 4 and 10 high-amplitude *b* and *y* ions (respectively) to theoretical ions of an ATPase peptide sequence from *S. enterica* (Figure 1A, B). The match had a high Ions score of 86.3. It is evident that the Ions score is directly related to the number of fragment ions matched because when matched *b* and *y* ion masses were iteratively deleted from the original spectrum peak list and the modified lists researched with Mascot, the Ions scores dropped (Figure 2). After the top 14 matched peaks were deleted, other, previously lower-amplitude *b* and *y* ions were then matched, but this resulted in a lower Ions score of 16. Deletion of another *b* ion mass led to an Ions score of 10 (Figure 1C), and after deletion of two more, Mascot could no longer match the depleted spectrum to the ATPase peptide, or any other peptide sequence candidate from the *S. enterica* protein database for that matter. Thus, the Mascot Ions score magnitude reflects the extent of identifiable *b*−*y* fragmentation. In essence, the low Ions score reflects poorer *b*−*y* ion assignment, and that is evidenced by the observation that the set of matched ions for the depleted spectrum in Figure 1C is really no better than the set for a false match with an Ions score of 8.2 between the original *S. enterica* spectrum and a nonhomologous human sequence (Figure 1D).

Note that when using this Orbitrap-LTQ configuration no benefit from the higher resolution Orbitrap is realized at the level of measuring fragment ions. Rather, the benefit is gained at another level with nothing to do with peptide fragmentation. Setting a wide ±1.5 Da PIM tolerance parameter window in Mascot for the spectrum in Figure 1A, as one might if the PIM were resolved in a linear ion trap, yields 296 tryptic peptide sequence candidates (one possible missed tryptic cleavage) from the *S. enterica* protein sequence database to be considered for matching. However, setting a narrow ±10 ppm PIM tolerance to reflect the greater accuracy of the Orbitrap yields only 7 tryptic peptide candidates. This means there are fewer candidates for matching in the more restrictive PIM search. This parametric change is reflected in the Mascot Identity score, a value directly

Received: October 4, 2010
Published: January 13, 2011

dx.doi.org/10.1021/pr101003r | *J. Proteome Res.* 2011, 10, 1432–1435

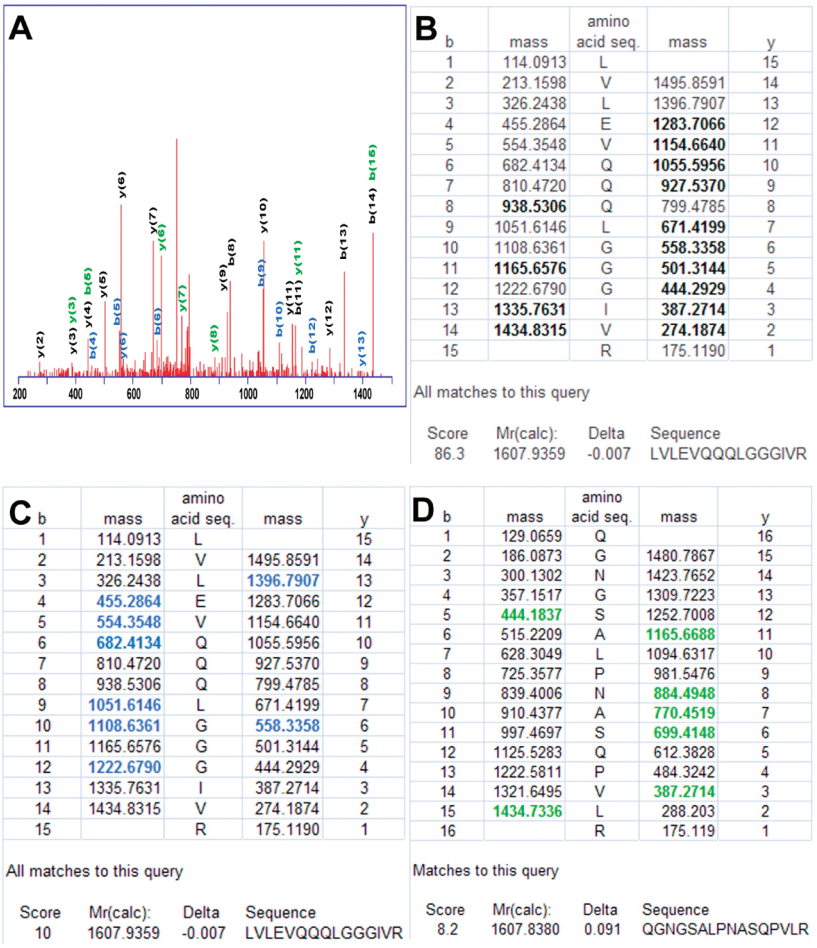


Figure 1. Mascot results of matched *b* and *y* ions from an LTQ-generated *S. enterica* spectrum with parent ion resolved in an Orbitrap. (A) Native *S. enterica* spectral peaks with matched *b* and *y* ions from (B) in black, (C) in blue and (D) in green. (B) Masses of 14 matched ions (bold black) between the *S. enterica* sequence LVLEVQQQLGGGIVR and the native spectrum producing an Ions score of 86.3. (C) Masses of 8 matched ions (bold blue) between the *S. enterica* sequence LVLEVQQQLGGGIVR and the native spectrum depleted of 15 *b* and *y* ions (14 of which were previously matched in (B)) producing an Ions score of 10. (D) Masses of 7 matched ions (bold green) between the human sequence QGNGSALPNASQPVLR and the native spectrum producing an Ions score of 8.2.

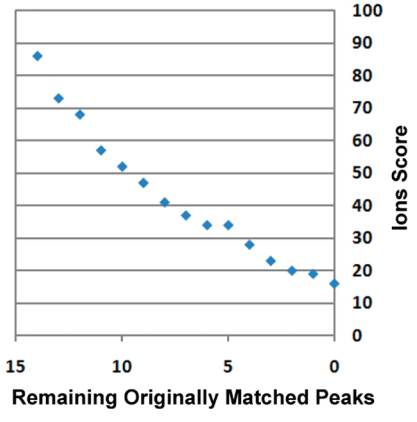


Figure 2. Plot of Mascot Ions score vs number of remaining matched peaks originally matched in spectrum in Figure 1A.

related to the number of peptide sequences from a given database with masses within the prescribed PIM tolerance window (peptide cleavage and mass modification parameters can also affect the Identity score). Consequently, the corresponding Identity score from the ±1.5 Da PIM search is 25 while the Identity score for the ±10 ppm PIM search is 8 (Table 1). This has important implications in context with Ions scores that will be discussed in the next paragraph. Meanwhile, note that the Ions score is unaffected by changing the PIM tolerance during the searches. For Figure 1B, a ± 10 ppm PIM tolerance was used. When a wider ±1.5 Da PIM search window was used instead, the top-ranked peptide-spectrum match was to the same ATPase peptide sequence, produced the same Ions score and identified the same *b*-*y* ions as in Figure 1B. Hence, the effect of using an Orbitrap to measure PIM more accurately is not improved fragment ion matching but reduced numbers of peptide sequence candidates considered.

To judge peptide-spectrum matches using Mascot scoring standards, one should consider the two independent Ions and Identity scores together. The Mascot-defined rule is that there is approximately a 95% chance (reported as expect = 0.05) that a spectrum-peptide sequence match is not random if the Ions score is equal to the Identity score.[6] In other words, if an Ions score is greater than the Identity score, then the confidence that the peptide-spectrum match is correct increases (expect decreases).

dx.doi.org/10.1021/pr101003r |*J. Proteome Res.* 2011, 10, 1432–1435

Table 1. Relation between the Number of Candidate Peptide Sequences from a Database with Masses within a Tolerance of the Parent Ion Mass[a] for Any Given Spectrum ("qmatch" in the Mascot results file) and the Calculated Mascot Identity Score [Identity score = 10 $\log_{10}$ qmatch]

candidates (qmatch)	Identity score
1	0
2	3
7	8
10	10
11	10
16	12
20	13
34	15
53	17
69	18
100	20
214	23
296	25
486	27
500	27
1000	30
2759	34
2776	34
4458	36
4524	37
10 000	40

[a] Other parameters also affect the number of candidates such as number of variable modifications, missed cleavages, consideration of ^{13}C peaks, and enzyme specificity, but these are held constant in this study.

Interestingly, one can easily lower the Identity score merely by using a more restrictive PIM tolerance parameter, making it seemingly justifiable under Mascot scoring rules to subsequently accept lower Ions scores. However, this may cause one to accept a "poorly fragmented" spectrum without a clearly discernible $b-y$ ion series merely because its low Ions score was greater than or equal to its low Identity score (e.g., Ions score of 10 vs Identity score of 8).

Obviously this is problematic, but not just because there is weak fragmentation evidence to support the peptide identification. As both scores approach zero, the Mascot rule for assessing match significance is no longer meaningful: a zero Ions score means the $b-y$ ion series cannot be reliably matched while a zero Identity score means there is only one candidate to which a match can be made. Therefore, when one accepts a spectrum with a low Ions score under restrictive PIM search tolerances that severely limit the number of candidates, the basis of identification shifts from fragment ion matching to PIM matching (recall that there is no benefit of Orbitrap PIM accuracy to fragment ion matching when using an LTQ for MS/MS). Under such conditions, the practice of accepting low Ions scores is anathema to the intended function of Mascot because the identification is no longer based on evidence of fragmentation but rather to candidates of similar mass from a constrained peptide sequence database. In reality, it would be incorrect to presume an observed PIM is best explained by a limited number of select peptide sequences from a genome sequence-derived database given the unknown number of molecules with similar mass that are biologically plausible in any system or inadvertently created during experimentation. For molecular identification by PIM alone, an unknown peptide similar to that denoted in Figure 1B would need to be resolved at better than 0.01 ppm to eliminate theoretical nonisomeric composition possibilities. By and large, this level of resolution was not achieved for most of the peptides in the studies in question, even after using postprocessing software that considers repeated measurements to lower PIM error estimates.[7]

These negative effects of basing peptide identification on PIM are perpetuated when trying to model FDRs. The most common method to model FDRs is to search a decoy database of the same size and relative composition as the true database and then apply an assumption that any match made to a decoy sequence is false. An Ions score cutoff can then be chosen that models for the true data the rate at which spectrum matching is false (i.e., a product of random $b-y$ ion matching[8]). Now, consider a protein sequence database with the amino acid sequences in reversed order (i.e., a "reversed" database, the easiest decoy to construct and the one most commonly used for FDR estimates). While reversing the sequences may create decoys, reversing also alters the masses of the hypothetical tryptic peptides derived from (nearly) any given protein. Consequently, a database of reversed sequences may lack an equivalent number of decoys. Table 2 shows that there can be inexact numbers of candidate peptides from different sized forward and reversed protein sequence databases when using ±1.5 Da and ±10 ppm PIM search tolerances for the spectrum in Figure 1A. Thus, these differences may lead to inaccurate modeling of FDRs.[9] Hence, it would be preferable to know *a priori* if decoys were devoid of systematic bias.[9]

Certainly, variability between decoy and true databases is tempered when greater numbers of candidates are examined.[10] This occurs when databases have a large number of records or if the PIM search tolerance window is wide (Table 2). However, under strict PIM search tolerances or when small databases are searched, the FDR modeling accuracy deteriorates as these numbers shrink.[10] Most problematic is when no false matches are made to sequences in the reversed database, a scenario exacerbated under restrictive PIM tolerance settings. When searching the spectrum in Figure 1A against the reversed *S. enterica* database using a ±10 ppm PIM tolerance, no matches were made (Table 2). Therefore, under these circumstances, there is no false match Ions score value for benchmarking. This could happen in any parallel situation where a database lacks a sufficient number of decoys or because PIM tolerance constraints severely limit the number of candidates for trial (Table 2). Either way, the number of false matches appears to be zero or some small number that leads to an artificially low FDR when calculated for the entire data set. Subsequently, the artificially low FDR drives down the Ions score cutoff that is imposed on the true data set as part of data quality standards. Thus, a result of undersampling is that spectra with low Ions scores are erroneously accepted. Furthermore, the FDR no longer accurately reflects the rate at which spectra are incorrectly identified by $b-y$ ion matching. Rather, the FDR is based on restricted PIM tolerances against a suboptimal decoy and may merely reflect a limited number of trials of a limited number of restricted masses.

In the papers in question, many peptides were identified with low Ions scores. Thus, it is possible that many of the spectra for the peptides with scores 16 or lower in these papers have no better $b-y$ ion matching than the depleted spectrum match in

Table 2. Number of Peptide Sequence Candidates in the Respective Forward and Reversed Protein Sequence Databases of the Following Organisms within ±1.5 Da and ±10 ppm Tolerance Ranges of the Parent Ion Mass (PIM) for the *Salmonella enterica* spectrum in Figure 1A, and the Ions Score for the Top-Ranked Match to a Candidate

	Salmonella enterica (4697 records)				*Saccharomyces cerevisiae* (6470 records)				*Homo sapiens* (34 352 records)				*Glycine max* (75 778 records)			
	forward		reverse		forward		reverse		forward		reverse		forward		reverse	
PIM tolerance	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score	candi-dates	top-ranked Ions score
1.5 Da	296	86	214	8[a]	500	8[a]	486	17[a]	2759	8[a]	2776	15[a]	4524	8[a]	4458	12[a]
10 ppm	7	86	2	no matches	11	no matches	16	no matches	34	0.17[a]	34	6[a]	53	no matches	69	0.12[a]

[a] Considered a false match.

Figure 1C or the false match in Figure 1D. Hence, there may be little $b-y$ ion fragmentation evidence that supports the amino acid order of these peptide sequence candidates or that supports the positioning of amino acid modifications. Furthermore, it appears that identification instead shifted to PIM matching, leading to inaccurate FDR modeling for the reasons stated. This could explain why peptides with low Ions scores were claimed.

It is important to stress that not all of the peptides in the papers in question are doubtful. Therefore, the aim of this letter is not to accuse, but to provoke: It may be possible to optimize decoy databases for more rigorous FDR modeling,[9] employ other measures that independently describe match confidence,[11,12] use broad PIM search tolerances to allow for more comparisons[13] and then restrict matches to peptides within a narrow mass window after the search,[14] measure fragment ions in the Orbitrap[15] and make use of the added level of mass discrimination to improve peptide identification confidence, or consider a number of other characteristics that validate data rather than restrict searches to just a few presumptive peptide candidates.

AUTHOR INFORMATION

Corresponding Author

*Bret Cooper, 10300 Baltimore Ave., Bldg. 006, Rm. 213, Beltsville, MD 20705. Phone, 301-504-9892; e-mail, bret.cooper@ars.usda.gov.

REFERENCES

(1) Soufi, B.; Kelstrup, C. D.; Stoehr, G.; Frohlich, F.; Walther, T. C.; Olsen, J. V. Global analysis of the yeast osmotic stress response by quantitative proteomics. *Mol. Biosyst.* **2009**, *5*, 1337–46.

(2) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **2009**, *325*, 834–40.

(3) Choudhary, C.; Olsen, J. V.; Brandts, C.; Cox, J.; Reddy, P. N.; Bohmer, F. D.; Gerke, V.; Schmidt-Arras, D. E.; Berdel, W. E.; Muller-Tidow, C.; Mann, M.; Serve, H. Mislocalized activation of oncogenic RTKs switches downstream signaling outcomes. *Mol. Cell* **2009**, *36*, 326–39.

(4) Gnad, F.; de Godoy, L. M.; Cox, J.; Neuhauser, N.; Ren, S.; Olsen, J. V.; Mann, M. High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* **2009**, *9*, 4642–52.

(5) Olsen, J. V.; Vermeulen, M.; Santamaria, A.; Kumar, C.; Miller, M. L.; Jensen, L. J.; Gnad, F.; Cox, J.; Jensen, T. S.; Nigg, E. A.; Brunak, S.; Mann, M. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **2010**, *3*, ra3.

(6) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–67.

(7) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–72.

(8) Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol. Cell. Proteomics* **2007**, *6*, 1599–608.

(9) Feng, J.; Naiman, D. Q.; Cooper, B. Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* **2007**, *23*, 2210–7.

(10) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell. Proteomics* **2008**, *7*, 962–70.

(11) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7*, 3354–63.

(12) Chen, Y.; Zhang, J.; Xing, G.; Zhao, Y. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. *J. Proteome Res.* **2009**, *8*, 3141–7.

(13) Ding, Y.; Choi, H.; Nesvizhskii, A. I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *J. Proteome Res.* **2008**, *7*, 4878–89.

(14) Hsieh, E. J.; Hoopmann, M. R.; MacLean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9*, 1138–43.

(15) Wenger, C. D.; McAlister, G. C.; Xia, Q.; Coon, J. J. Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2010**, *9*, 754–63.